



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2023-0076325
(43) 공개일자 2023년05월31일

(51) 국제특허분류(Int. Cl.)
H04N 21/2343 (2011.01) H04N 21/2187 (2011.01)
H04N 21/24 (2011.01) H04N 21/478 (2011.01)
(52) CPC특허분류
H04N 21/234309 (2013.01)
H04N 21/2187 (2013.01)
(21) 출원번호 10-2021-0163149
(22) 출원일자 2021년11월24일
심사청구일자 2021년11월24일

(71) 출원인
인하대학교 산학협력단
인천광역시 미추홀구 인하로 100(용현동, 인하대학교)
(72) 발명자
송민석
서울특별시 서초구 남부순환로 2311-12, 107동 402호(방배동, 래미안방배아트힐)
이다영
인천광역시 미추홀구 인하로119번길 7, 406호(용현동, 청호주택)
(74) 대리인
양성보

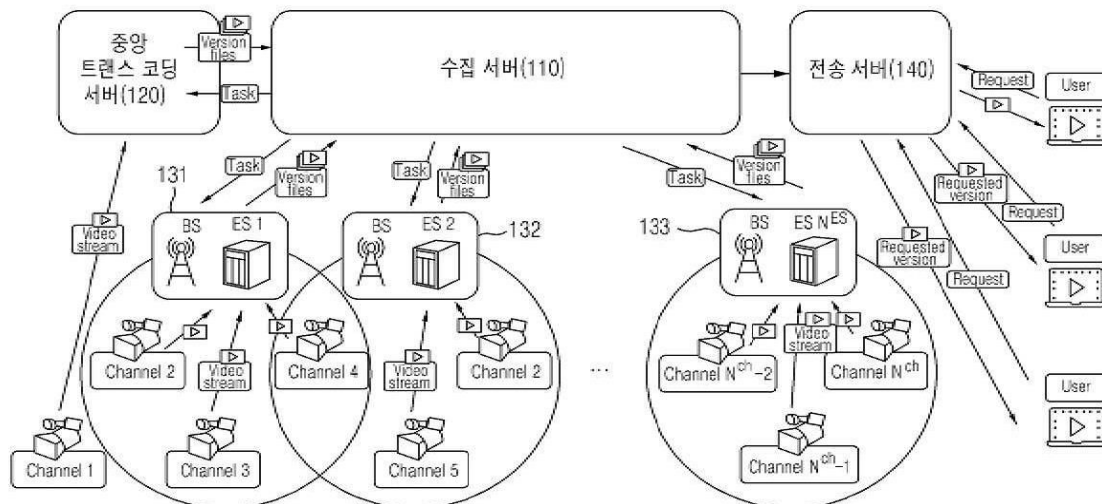
전체 청구항 수 : 총 8 항

(54) 발명의 명칭 옛지 컴퓨팅 환경에서 라이브 비디오 스트리밍 품질을 고려한 비용 최적화를 위한 트랜스코딩 작업 할당 방법 및 시스템

(57) 요약

옛지 컴퓨팅 환경에서 라이브 비디오 스트리밍 품질을 고려한 비용 최적화를 위한 트랜스코딩 작업 할당 방법 및 시스템이 제시된다. 본 발명에서 제안하는 옛지 컴퓨팅 환경에서 라이브 비디오 스트리밍 품질을 고려한 비용 최적화를 위한 트랜스코딩 작업 할당 방법은 수집 서버를 통해 옛지 컴퓨팅 환경에서 인기도 가중 비디오 품질 대 (뒷면에 계속)

대표도



비 비용 효율이 높은 순서대로 중앙 트랜스코딩 서버에서 수행되는 트랜스코딩 작업에 우선 순위를 부여하는 단계, 상기 수집 서버를 통해 상기 우선 순위가 부여된 각각의 트랜스코딩 작업에 대하여 네트워크 범위 및 계산 부하에 따라 해당 트랜스코딩 작업을 중앙 트랜스코딩 서버 또는 엣지 서버에 할당하는 단계, 상기 인기도 가중 비디오 품질에 대한 네트워크 범위 및 계산 부하에 따른 영향을 최소화하기 위하여 상기 수집 서버를 통해 일부 작업을 라이브 스트리밍 공급자가 운영하는 상기 중앙 트랜스코딩 서버에 재할당하거나 또는 제거하는 단계 및 상기 수집 서버를 통해 할당된 트랜스코딩 작업을 상기 중앙 트랜스코딩 서버 또는 상기 엣지 서버에서 수행한 후, 전송 서버를 통해 복수의 시청자 단말에 라이브 비디오 스트리밍을 제공하는 단계를 포함한다.

(52) CPC특허분류

H04N 21/23439 (2013.01)

H04N 21/2402 (2013.01)

H04N 21/4781 (2013.01)

이 발명을 지원한 국가연구개발사업

과제고유번호	1711135755
과제번호	NRF-2021R1F1A1064053
부처명	과학기술정보통신부
과제관리(전문)기관명	한국연구재단
연구사업명	이공분야기초연구사업, 기본연구
연구과제명	라이브 스트리밍 환경에서 스트리밍 품질 향상을 위한 포그/엣지 협업 기반 최적 위
크로드 할당 기법에 대한 연구	
기 여 율	1/1
과제수행기관명	인하대학교
연구기간	2021.06.01 ~ 2022.05.31

명세서

청구범위

청구항 1

수집 서버를 통해 엣지 컴퓨팅 환경에서 인기도 가중 비디오 품질(Popularity-Weighted Video Quality; PWQ) 대비 비용 효율이 높은 순서대로 중앙 트랜스코딩 서버에서 수행되는 트랜스코딩 작업에 우선 순위를 부여하는 단계;

상기 수집 서버를 통해 상기 우선 순위가 부여된 각각의 트랜스코딩 작업에 대하여 네트워크 범위 및 계산 부하에 따라 해당 트랜스코딩 작업을 중앙 트랜스코딩 서버 또는 엣지 서버에 할당하는 단계;

상기 인기도 가중 비디오 품질에 대한 네트워크 범위 및 계산 부하에 따른 영향을 최소화하기 위하여 상기 수집 서버를 통해 일부 작업을 라이브 스트리밍 공급자가 운영하는 상기 중앙 트랜스코딩 서버에 재할당하거나 또는 제거하는 단계; 및

상기 수집 서버를 통해 할당된 트랜스코딩 작업을 상기 중앙 트랜스코딩 서버 또는 상기 엣지 서버에서 수행한 후, 전송 서버를 통해 복수의 시청자 단말에 라이브 비디오 스트리밍을 제공하는 단계

를 포함하는 트랜스코딩 작업 할당 방법.

청구항 2

제1항에 있어서,

상기 수집 서버를 통해 엣지 컴퓨팅 환경에서 인기도 가중 비디오 품질 대비 비용 효율이 높은 순서대로 중앙 트랜스코딩 서버에서 수행되는 트랜스코딩 작업에 우선 순위를 부여하는 단계는,

상기 수집 서버에서 채널의 목록이 존재하고, 각 채널의 비디오 스트림을 캡처 및 인코딩하여 업로드하며,

복수의 시청자 단말로부터 요청된 비트레이트를 가진 버전의 비디오 스트림을 이진 값의 조합으로 정의하고 인기도 가중 비디오 품질을 계산하며, 프로세싱 용량 제약 및 네트워크 범위 제약 내에서 복수의 사용자 단말에게 전송되는 비디오 품질을 최대화 하기 위한 인기도 가중 비디오 품질에 따라 트랜스코딩 작업을 할당하며, 이후 비용 예산 제약이 충족될 때까지 인기도 가중 비디오 품질 손실을 최소화하도록 상기 할당된 작업을 제거하는

트랜스코딩 작업 할당 방법.

청구항 3

제2항에 있어서,

상기 수집 서버를 통해 상기 우선 순위가 부여된 각각의 트랜스코딩 작업에 대하여 네트워크 범위 및 계산 부하에 따라 해당 트랜스코딩 작업을 중앙 트랜스코딩 서버 또는 엣지 서버에 할당하는 단계는,

각 채널이 트랜스코딩 서버의 네트워크 범위 내에 있는지 여부를 나타내는 이진 변수를 사용하여, 해당 채널의 가장 낮은 비트레이트 버전에 대한 트랜스코딩 작업 실행 여부에 따라 이진값을 부여하는

트랜스코딩 작업 할당 방법.

청구항 4

제3항에 있어서,

상기 상기 수집 서버를 통해 할당된 트랜스코딩 작업을 상기 중앙 트랜스코딩 서버 또는 상기 엣지 서버에서 수행한 후, 전송 서버를 통해 복수의 시청자 단말에 라이브 비디오 스트리밍을 제공하는 단계는,

복수의 시청자 단말로부터 요청된 비트레이트를 가진 버전이 트랜스코딩 작업이 실행되지 않은 경우, 트랜스코딩 작업이 실행된 버전 중에서 상기 요청된 비트레이트보다 낮은 버전 중 가장 높은 비트레이트를 가진 버전을 전송하는

트랜스코딩 작업 할당 방법.

청구항 5

엣지 컴퓨팅 환경에서 인기도 가중 비디오 품질(Popularity-Weighted Video Quality; PWQ) 대비 비용 효율이 높은 순서대로 중앙 트랜스코딩 서버에서 수행되는 트랜스코딩 작업에 우선 순위를 부여하는 수집 서버;

상기 수집 서버를 통해 상기 우선 순위가 부여된 각각의 트랜스코딩 작업에 대하여 네트워크 범위 및 계산 부하에 따라 해당 트랜스코딩 작업을 할당 받는 중앙 트랜스코딩 서버 및 엣지 서버 -상기 중앙 트랜스코딩 서버는 상기 인기도 가중 비디오 품질에 대한 네트워크 범위 및 계산 부하에 따른 영향을 최소화하기 위하여 상기 수집 서버를 통해 일부 작업을 재할당 받거나 또는 일부 작업이 제거됨-; 및

상기 수집 서버를 통해 할당된 트랜스코딩 작업을 상기 중앙 트랜스코딩 서버 또는 상기 엣지 서버에서 수행한 후, 복수의 시청자 단말에 라이브 비디오 스트리밍을 제공하는 전송 서버

를 포함하는 트랜스코딩 작업 할당 시스템.

청구항 6

제5항에 있어서,

상기 수집 서버는,

채널의 목록이 존재하고, 각 채널의 비디오 스트림을 캡처 및 인코딩하여 업로드하며,

복수의 시청자 단말로부터 요청된 비트레이트를 가진 버전의 비디오 스트림을 이진 값의 조합으로 정의하고 인기도 가중 비디오 품질을 계산하며, 프로세싱 용량 제약 및 네트워크 범위 제약 내에서 복수의 사용자 단말에게 전송되는 비디오 품질을 최대화 하기 위한 인기도 가중 비디오 품질에 따라 트랜스코딩 작업을 할당하며, 이후 비용 예산 제약이 충족될 때까지 인기도 가중 비디오 품질 손실을 최소화하도록 상기 할당된 작업을 제거하는

트랜스코딩 작업 할당 시스템.

청구항 7

제6항에 있어서,

상기 중앙 트랜스코딩 서버 및 엣지 서버는,

각 채널이 트랜스코딩 서버의 네트워크 범위 내에 있는지 여부를 나타내는 이진 변수를 사용하여, 해당 채널의 가장 낮은 비트레이트 버전에 대한 트랜스코딩 작업 실행 여부에 따라 이진값을 부여 받는

트랜스코딩 작업 할당 시스템.

청구항 8

제3항에 있어서,

상기 전송 서버는,

복수의 시청자 단말로부터 요청된 비트레이트를 가진 버전이 트랜스코딩 작업이 실행되지 않은 경우, 트랜스코딩 작업이 실행된 버전 중에서 상기 요청된 비트레이트보다 낮은 버전 중 가장 높은 비트레이트를 가진 버전을 전송하는

트랜스코딩 작업 할당 시스템.

발명의 설명

기술 분야

본 발명은 엣지 컴퓨팅 환경에서 라이브 비디오 스트리밍 품질을 고려한 비용 최적화를 위한 트랜스코딩 작업 할당 방법 및 시스템에 관한 것이다.

배경 기술

[0001]

- [0002] 비디오 스포츠 경기와 같은 실시간 콘텐츠의 라이브 스트리밍이 인기를 얻고 있다. 예를 들어, 최고의 게임 생중계 플랫폼 중 하나인 Twitch는 미국 인터넷 트래픽의 1.8%를 차지한다. 라이브 스트리밍 시나리오에서 방송사의 채널은 비디오 스트림을 지속적으로 캡처하고 인코딩하여 지역 서버에 업로드한 다음 요청한 시청자에게 전달한다. 이러한 지속적인 컴퓨팅 및 전송 프로세스에는 많은 컴퓨팅 리소스가 필요하다.
- [0003] 변화하는 네트워크 조건을 허용하기 위해 라이브 스트리밍 서비스는 일반적으로 HTTP(DASH)를 통한 동적 적응 스트리밍을 사용한다. 여기서 각 소스 비디오 스트림은 다른 비트레이트의 버전으로 트랜스코딩되어 시청자에게 전달되는 비디오가 네트워크 대역폭과 일치한다. 그러나 트랜스코딩에는 단일 지역 서버에서 수행할 수 있는 것보다 더 많은 계산이 필요하므로 방송사는 제한된 수의 비디오 스트림 버전만 제공한다. 예를 들어, Twitch.tv는 프리미엄 채널에만 트랜스코딩 서비스를 제공한다. 따라서 많은 시청자가 최적의 아닌 비트레이트로 비디오를 수신하여 시청자 QoE를 저하시킨다.
- [0004] 이를 위해 서비스 제공자는 클라우드 서비스를 사용할 수 있지만, 클라우드 서버가 방송 소스에서 멀리 떨어져 있는 경우 라이브 스트리밍의 대기 시간 요구 사항을 충족하지 못한다. 그로 인해 등장한 엣지 컴퓨팅 환경에서는 처리할 데이터 근처에 엣지 서버(Edge Server; ES)를 배치한다. Gartner는 2023년까지 대기업의 50% 이상이 엣지 컴퓨팅을 필요로 하는 6개 이상의 사물 인터넷(Internet of Things; IoT) 애플리케이션을 보유할 것이며 엣지 인프라의 수가 증가할 것으로 예측한다.
- [0005] 라이브 스트리밍 서비스 제공업체는 MEC를 사용하여 채널 위치에 더 가까운 엣지 서버에서 비디오 스트림을 트랜스코딩함으로써 이전에 중앙 수집 서버로 전달되었던 트랜스코딩 워크로드를 분산할 수 있다. 이는 스트리밍 데이터 경로에 있기 때문에 대기 시간이 거의 증가하지 않으나, 각 엣지 서버의 CPU 프로세싱 용량과 네트워크 범위는 제한되어 있으므로 트랜스코딩 작업의 효과적인 할당이 필요하다. 또한 엣지 인프라를 사용하는 총 비용을 최소화할 수 있어야 한다.
- [0006] 엣지 컴퓨팅을 사용한 비디오 스트리밍에서의 계산 부하를 줄이는 것을 목표로 한 기존의 기법들은 VoD 시스템에 적용된 것이 대부분이며, 지속적인 실시간 트랜스코딩이 필요한 라이브 스트리밍 시스템에 대한 기여가 부족하다. 라이브 스트리밍 시스템에서 일부 트랜스코딩 작업에 시청자의 장치를 사용하는 몇 가지 기술이 제안되었으나, 시청자 연결의 불안정성으로 인해 신뢰성 높은 서비스를 제공하기 어렵다.

선행기술문헌

특허문헌

- [0007] (특허문헌 0001) 한국 등록특허공보 제10-1569502호(2015.11.10)

발명의 내용

해결하려는 과제

- [0008] 본 발명이 이루고자 하는 기술적 과제는 엣지 컴퓨팅 환경의 라이브 스트리밍 서버의 엣지 서버의 사용률을 제어함으로써 비용 제약 조건에서 총 인기도 가중 비디오 품질을 최대화하기 위한 방법 및 시스템을 제공하는데 있다. 엣지 서버 컴퓨팅을 사용하여 라이브 스트리밍 시스템에서의 트랜스코딩을 위해 서비스 제공자의 비용 한도 제약, 각 엣지 서버의 프로세싱 용량 및 네트워크 범위 제약 내에서 사용자 단말에게 전송되는 종합적인 비디오 품질을 극대화하기 위해 각 채널에 필요한 비트레이트 버전을 선택하고, 이를 생성하는 트랜스코딩 작업을 각 엣지 서버에 할당한다.

과제의 해결 수단

- [0009] 일 측면에 있어서, 본 발명에서 제안하는 엣지 컴퓨팅 환경에서 라이브 비디오 스트리밍 품질을 고려한 비용 최적화를 위한 트랜스코딩 작업 할당 방법은 수집 서버를 통해 엣지 컴퓨팅 환경에서 인기도 가중 비디오 품질(Popularity-Weighted Video Quality; PWQ) 대비 비용 효율이 높은 순서대로 중앙 트랜스코딩 서버에서 수행되는 트랜스코딩 작업에 우선 순위를 부여하는 단계, 상기 수집 서버를 통해 상기 우선 순위가 부여된 각각의 트랜스코딩 작업에 대하여 네트워크 범위 및 계산 부하에 따라 해당 트랜스코딩 작업을 중앙 트랜스코딩 서버 또는 엣지 서버에 할당하는 단계, 상기 인기도 가중 비디오 품질에 대한 네트워크 범위 및 계산 부하에 따른 영향

을 최소화하기 위하여 상기 수집 서버를 통해 일부 작업을 라이브 스트리밍 공급자가 운영하는 상기 중앙 트랜스코딩 서버에 재할당하거나 또는 제거하는 단계 및 상기 수집 서버를 통해 할당된 트랜스코딩 작업을 상기 중앙 트랜스코딩 서버 또는 상기 엣지 서버에서 수행한 후, 전송 서버를 통해 복수의 시청자 단말에 라이브 비디오 스트리밍을 제공하는 단계를 포함한다.

[0010] 상기 수집 서버를 통해 엣지 컴퓨팅 환경에서 인기도 가중 비디오 품질 대비 비용 효율이 높은 순서대로 중앙 트랜스코딩 서버에서 수행되는 트랜스코딩 작업에 우선 순위를 부여하는 단계는 상기 수집 서버에서 채널의 목록이 존재하고, 각 채널의 비디오 스트림을 캡처 및 인코딩하여 업로드하며, 복수의 시청자 단말로부터 요청된 비트레이트를 가진 버전의 비디오 스트림을 이진 값의 조합으로 정의하고 인기도 가중 비디오 품질을 계산하며, 프로세싱 용량 제약 및 네트워크 범위 제약 내에서 복수의 사용자 단말에게 전송되는 비디오 품질을 최대화 하기 위한 인기도 가중 비디오 품질에 따라 트랜스코딩 작업을 할당하며, 이후 비용 예산 제약이 충족될 때까지 인기도 가중 비디오 품질 손실을 최소화하도록 상기 할당된 작업을 제거한다.

[0011] 상기 수집 서버를 통해 상기 우선 순위가 부여된 각각의 트랜스코딩 작업에 대하여 네트워크 범위 및 계산 부하에 따라 해당 트랜스코딩 작업을 중앙 트랜스코딩 서버 또는 엣지 서버에 할당하는 단계는 각 채널이 트랜스코딩 서버의 네트워크 범위 내에 있는지 여부를 나타내는 이진 변수를 사용하여, 해당 채널의 가장 낮은 비트레이트 버전에 대한 트랜스코딩 작업 실행 여부에 따라 이진값을 부여한다.

[0012] 상기 상기 수집 서버를 통해 할당된 트랜스코딩 작업을 상기 중앙 트랜스코딩 서버 또는 상기 엣지 서버에서 수행한 후, 전송 서버를 통해 복수의 시청자 단말에 라이브 비디오 스트리밍을 제공하는 단계는 복수의 시청자 단말로부터 요청된 비트레이트를 가진 버전이 트랜스코딩 작업이 실행되지 않은 경우, 트랜스코딩 작업이 실행된 버전 중에서 상기 요청된 비트레이트보다 낮은 버전 중 가장 높은 비트레이트를 가진 버전을 전송한다.

[0013] 또 다른 일 측면에 있어서, 본 발명에서 제안하는 엣지 컴퓨팅 환경에서 라이브 비디오 스트리밍 품질을 고려한 비용 최적화를 위한 트랜스코딩 작업 할당 시스템은 엣지 컴퓨팅 환경에서 인기도 가중 비디오 품질 (Popularity-Weighted Video Quality; PWQ) 대비 비용 효율이 높은 순서대로 중앙 트랜스코딩 서버에서 수행되는 트랜스코딩 작업에 우선 순위를 부여하는 수집 서버, 상기 수집 서버를 통해 상기 우선 순위가 부여된 각각의 트랜스코딩 작업에 대하여 네트워크 범위 및 계산 부하에 따라 해당 트랜스코딩 작업을 할당 받는 중앙 트랜스코딩 서버 및 엣지 서버 -상기 중앙 트랜스코딩 서버는 상기 인기도 가중 비디오 품질에 대한 네트워크 범위 및 계산 부하에 따른 영향을 최소화하기 위하여 상기 수집 서버를 통해 일부 작업을 재할당 받거나 또는 일부 작업이 제거됨- 및 상기 수집 서버를 통해 할당된 트랜스코딩 작업을 상기 중앙 트랜스코딩 서버 또는 상기 엣지 서버에서 수행한 후, 복수의 시청자 단말에 라이브 비디오 스트리밍을 제공하는 전송 서버를 포함한다.

발명의 효과

[0014] 본 발명의 실시예들에 따르면 다중 액세스 엣지 컴퓨팅(Multiple-access Edge-Computing; MEC) 아키텍처를 사용하는 라이브 비디오 스트리밍 시스템(Live Video Streaming System)을 통해 각 엣지 서버의 CPU 프로세싱 용량, 무선 네트워크 범위, 방송사의 비용 예산 및 시청자 비디오 품질을 고려하여 각 엣지 서버에 트랜스코딩 작업을 할당할 수 있다. 또한, 본 발명의 실시예에 따르면, 동일한 비용 예산에서 종래기술에 따른 기법보다 제안하는 기법이 평균 1.4배 더 많은 PWQ를 달성할 수 있고, MEC 환경에서 라이브 스트리밍 시스템의 효과적인 비용 관리에 기여할 수 있다.

도면의 간단한 설명

[0015] 도 1은 본 발명의 일 실시예에 따른 엣지 컴퓨팅 환경에서 라이브 비디오 스트리밍 품질을 고려한 비용 최적화를 위한 트랜스코딩 작업 할당 시스템의 구성을 나타내는 도면이다.

도 2는 본 발명의 일 실시예에 따른 엣지 컴퓨팅 환경에서 라이브 비디오 스트리밍 품질을 고려한 비용 최적화를 위한 트랜스코딩 작업 할당 방법을 설명하기 위한 흐름도이다.

도 3은 본 발명의 일 실시예에 따른 트랜스코딩 작업 결정 및 엣지 할당 알고리즘을 나타내는 도면이다.

도 4는 본 발명의 일 실시예에 따른 비용 절감을 위한 트랜스코딩 작업 재할당 및 제거 알고리즘을 나타내는 도면이다.

도 5는 본 발명의 일 실시예에 따른 트랜스코딩 작업 할당 기법과 종래기술의 기법과의 비용 제한에 따른 인기도 가중 품질을 비교한 도면이다.

도 6은 본 발명의 일 실시예에 따른 트랜스코딩 작업 할당 기법과 종래기술의 기법과의 채널 수에 따른 인기도 가중 품질을 비교한 도면이다.

도 7은 본 발명의 일 실시예에 따른 트랜스코딩 작업 할당 기법과 종래기술의 기법과의 엣지 서버 수에 따른 인기도 가중 품질을 비교한 도면이다.

도 8은 본 발명의 일 실시예에 따른 트랜스코딩 작업 할당 기법과 종래기술의 기법과의 비용 제한에 따른 온-오프(On-off) 비용 모델에서의 인기도 가중 품질을 비교한 도면이다.

도 9는 본 발명의 일 실시예에 따른 트랜스코딩 작업 할당 기법의 인기도 가중 품질을 선형 비용 모델과 온-오프(On-off) 비용 모델에서 비교한 도면이다.

발명을 실시하기 위한 구체적인 내용

- [0016] 본 발명은 다중 액세스 엣지 컴퓨팅(Multiple-access Edge-Computing; MEC) 아키텍처를 사용하는 라이브 비디오 스트리밍 시스템(Live Video Streaming System)에서 각 엣지 서버의 CPU 프로세싱 용량, 무선 네트워크 범위, 방송사의 비용 예산 및 시청자 비디오 품질을 고려하여 각 엣지 서버에 트랜스코딩 작업을 할당하는 방법 및 시스템에 관한 것이다. 본 발명은 엣지 컴퓨팅 환경에서의 라이브 스트리밍을 수행하는 시스템에서 각 엣지 서버의 사용률 제어를 통해 비용 문제를 해결하는 방법 및 시스템에 관한 것이다.
- [0017] 본 발명은 먼저 인기도 가중 비디오 품질(Popularity-Weighted Video Quality; PWQ) 대비 비용 효율이 더 높은 트랜스코딩 작업에 우선 순위를 부여하고, 각 작업을 네트워크 범위와 계산 부하를 고려하여 엣지 서버에 할당한다. 그 후, 전체 PWQ에 대한 이러한 변경의 영향을 최소화하면서 일부 작업을 라이브 스트리밍 공급자가 운영하는 중앙 트랜스코딩 서버에 재할당하거나 제거하여 비용 예산을 충족한다.
- [0018] 이와 같이, 스트리밍 되는 비디오 품질과 엣지 서버의 채널 네트워크를 고려하여 트랜스코딩 작업을 엣지 서버에 할당함으로써 서비스 제공자의 비용 한도를 최적화하는 휴리스틱 알고리즘을 제안한다.
- [0019] 본 발명의 실시예에 따른 시뮬레이션 결과에 따르면, 동일한 비용 예산에서 종래기술에 따른 기법보다 제안하는 기법이 평균 1.4배 더 많은 PWQ를 달성한다. 이 차이는 비용 예산 여유가 적을 수록, 많은 채널이 있을 수록, 그리고 엣지 서버 수가 적을 수록 더욱 효과적이다.
- [0020] 또한, 본 발명의 실시예에 따른 실험 결과에 따르면, 제안하는 기법이 MEC 환경에서 라이브 스트리밍 시스템의 효과적인 비용 관리에 기여할 수 있음을 보인다. 이하, 본 발명의 실시 예를 첨부된 도면을 참조하여 상세하게 설명한다.
- [0022] 도 1은 본 발명의 일 실시예에 따른 엣지 컴퓨팅 환경에서 라이브 비디오 스트리밍 품질을 고려한 비용 최적화를 위한 트랜스코딩 작업 할당 시스템의 구성을 나타내는 도면이다.
- [0023] 제안하는 엣지 컴퓨팅 환경에서 라이브 비디오 스트리밍 품질을 고려한 비용 최적화를 위한 트랜스코딩 작업 할당 시스템은 수집 서버(Ingest server)(110), 중앙 트랜스코딩 서버(Central transcoding server; CTS)(120), 엣지 서버(Edge server; ES)(131, 132, ..., 133) 및 전송 서버(Delivery server)(140)를 포함한다.
- [0024] 본 발명의 실시예에 따른 수집 서버(110)는 엣지 컴퓨팅 환경에서 인기도 가중 비디오 품질(Popularity-Weighted Video Quality; PWQ) 대비 비용 효율이 높은 순서대로 중앙 트랜스코딩 서버에서 수행되는 트랜스코딩 작업에 우선 순위를 부여한다.
- [0025] 본 발명의 실시예에 따른 수집 서버(110)에는 채널의 목록이 존재하고, 각 채널의 비디오 스트림을 캡처 및 인코딩하여 업로드한다.
- [0026] 본 발명의 실시예에 따른 수집 서버(110)는 복수의 시청자 단말(User)로부터 요청된 비트레이트를 가진 버전의 비디오 스트림을 이진 값의 조합으로 정의하고 인기도 가중 비디오 품질을 계산한다. 그리고, 프로세싱 용량 제약 및 네트워크 범위 제약 내에서 복수의 사용자 단말에게 전송되는 비디오 품질을 최대화 하기 위한 인기도 가중 비디오 품질에 따라 트랜스코딩 작업을 할당하며, 이후 비용 예산 제약이 충족될 때까지 인기도 가중 비디오 품질 손실을 최소화하도록 상기 할당된 작업을 제거한다.
- [0027] 본 발명의 실시예에 따른 중앙 트랜스코딩 서버(120) 및 엣지 서버(131, 132, ..., 133)는 수집 서버를 통해 우선 순위가 부여된 각각의 트랜스코딩 작업에 대하여 네트워크 범위 및 계산 부하에 따라 해당 트랜스코딩 작업

을 할당 받는다.

- [0028] 중앙 트랜스코딩 서버(120)는 인기도 가중 비디오 품질에 대한 네트워크 범위 및 계산 부하에 따른 영향을 최소화하기 위하여 수집 서버(110)를 통해 일부 작업을 재할당 받거나 또는 일부 작업이 제거될 수 있다.
- [0029] 본 발명의 실시예에 따른 중앙 트랜스코딩 서버(120) 및 엣지 서버(131, 132, ..., 133)는 각 채널이 트랜스코딩 서버의 네트워크 범위 내에 있는지 여부를 나타내는 이진 변수를 사용하여, 해당 채널의 가장 낮은 비트레이트 버전에 대한 트랜스코딩 작업 실행 여부에 따라 이진값을 부여 받는다.
- [0030] 본 발명의 실시예에 따른 전송 서버(140)는 수집 서버(110)를 통해 할당된 트랜스코딩 작업을 중앙 트랜스코딩 서버(120) 또는 엣지 서버(131, 132, ..., 133)에서 수행한 후, 복수의 시청자 단말(User)에 라이브 비디오 스트리밍을 제공한다.
- [0031] 본 발명의 실시예에 따른 전송 서버(140)는 복수의 시청자 단말(User)로부터 요청된 비트레이트를 가진 버전이 트랜스코딩 작업이 실행되지 않은 경우, 트랜스코딩 작업이 실행된 버전 중에서 상기 요청된 비트레이트보다 낮은 버전 중 가장 높은 비트레이트를 가진 버전을 전송한다.
- [0032] 도 1을 참조하여, 제안하는 엣지 컴퓨팅 환경에서 라이브 비디오 스트리밍 품질을 고려한 비용 최적화를 위한 트랜스코딩 작업 할당 시스템에 대하여 더욱 상세히 설명한다.
- [0033] 본 발명의 실시예에 따르면, 각 채널에서 방송사의 비디오 소스가 송출되며, 이 채널들은 수집 서버(110)에서 관리된다. 수집 서버(110)에서는 채널의 목록이 존재하며, 각 채널의 비디오 스트림을 지속적으로 캡처, 인코딩하여 업로드된다, 각 채널의 원본 비디오 스트림은 라이브 스트리밍 공급자가 운영하는 중앙 트랜스코딩 서버(120)에 의해 트랜스코딩될 수 있다. 또한 채널에서 수집 서버(110)로의 업로드 경로가 복수의 엣지 서버(131, 132, ..., 133)를 통과하거나 이에 근접하며, 이러한 엣지 서버(131, 132, ..., 133)를 트랜스코딩에 사용할 수 있다고 가정한다.
- [0034] 수집 서버(110)는 각 트랜스코딩 작업을 실행할 서버를 결정한다. 그런 다음 전달 서버(140)는 복수의 시청자 단말에 적절한 비트레이트로 비디오 버전을 제공할 수 있다.
- [0035] N^{ES} 는 엣지 서버의 수이고, 트랜스코딩 서버(Transcoding Server; TS)의 총 수는 $N^{ES}+1$ 이다. 여기서 TS 0은 CTS(Central transcoding server)이기 때문이다. 다른 TS는 다른 CPU 프로세싱 용량을 가지므로, 본 발명에서는 TS j 의 CPU 프로세싱 용량 U_j^{max} ($j=0, \dots, N^{ES}$)를 CPU 프로세싱 능력이 가장 가장 높은 TS에 대비한 비율로 표현한다.
- [0036] 각 엣지 서버는 네트워크 범위 내에서 채널의 트랜스코딩 작업을 처리하기 위해 기지국에 할당된다. 따라서 각 엣지 서버는 채널로부터의 거리에 따라 네트워크 범위 제약이 있다. $C_{i,j}^{net}$ 는 각 채널 i 가 TS j 의 네트워크 범위 내에 있는지 여부를 나타내는 이진 변수이다. i 채널이 TS j 의 네트워크 범위 내에 있으면 $C_{i,j}^{net}=1$, 그렇지 않으면 $C_{i,j}^{net}=0$ 이다. 모든 트랜스코딩 작업은 CTS(TS 0)에서 실행할 수 있으므로 $\forall i, C_{i,0}^{net}=1$ 이다.
- [0038] 도 2는 본 발명의 일 실시예에 따른 엣지 컴퓨팅 환경에서 라이브 비디오 스트리밍 품질을 고려한 비용 최적화를 위한 트랜스코딩 작업 할당 방법을 설명하기 위한 흐름도이다.
- [0039] 제안하는 제안하는 엣지 컴퓨팅 환경에서 라이브 비디오 스트리밍 품질을 고려한 비용 최적화를 위한 트랜스코딩 작업 할당 방법은 수집 서버를 통해 엣지 컴퓨팅 환경에서 인기도 가중 비디오 품질(Popularity-Weighted Video Quality; PWQ) 대비 비용 효율이 높은 순서대로 중앙 트랜스코딩 서버에서 수행되는 트랜스코딩 작업에 우선 순위를 부여하는 단계(210), 상기 수집 서버를 통해 상기 우선 순위가 부여된 각각의 트랜스코딩 작업에 대하여 네트워크 범위 및 계산 부하에 따라 해당 트랜스코딩 작업을 중앙 트랜스코딩 서버 또는 엣지 서버에 할당하는 단계(220), 상기 인기도 가중 비디오 품질에 대한 네트워크 범위 및 계산 부하에 따른 영향을 최소화하기 위하여 상기 수집 서버를 통해 일부 작업을 라이브 스트리밍 공급자가 운영하는 상기 중앙 트랜스코딩 서버에 재할당하거나 또는 제거하는 단계(230) 및 상기 수집 서버를 통해 할당된 트랜스코딩 작업을 상기 중앙 트랜스코딩 서버 또는 상기 엣지 서버에서 수행한 후, 전송 서버를 통해 복수의 시청자 단말에 라이브 비디오 스트리밍을 제공하는 단계(240)를 포함한다.

- [0040] 단계(210)에서, 수집 서버를 통해 엣지 컴퓨팅 환경에서 인기도 가중 비디오 품질 대비 비용 효율이 높은 순서대로 중앙 트랜스코딩 서버에서 수행되는 트랜스코딩 작업에 우선 순위를 부여한다.
- [0041] 본 발명의 실시예에 따른 수집 서버에는 채널의 목록이 존재하고, 각 채널의 비디오 스트림을 캡처 및 인코딩하여 업로드한다.
- [0042] 본 발명의 실시예에 따른 수집 서버는 복수의 시청자 단말로부터 요청된 비트레이트를 가진 버전의 비디오 스트림을 이진 값의 조합으로 정의하고 인기도 가중 비디오 품질을 계산한다. 그리고, 프로세싱 용량 제약 및 네트워크 범위 제약 내에서 복수의 사용자 단말에게 전송되는 비디오 품질을 최대화 하기 위한 인기도 가중 비디오 품질에 따라 트랜스코딩 작업을 할당하며, 이후 비용 예산 제약이 충족될 때까지 인기도 가중 비디오 품질 손실을 최소화하도록 상기 할당된 작업을 제거한다.
- [0043] 단계(220)에서, 수집 서버를 통해 상기 우선 순위가 부여된 각각의 트랜스코딩 작업에 대하여 네트워크 범위 및 계산 부하에 따라 해당 트랜스코딩 작업을 중앙 트랜스코딩 서버 또는 엣지 서버에 할당한다.
- [0044] 본 발명의 실시예에 따른 중앙 트랜스코딩 서버 및 엣지 서버는 수집 서버를 통해 우선 순위가 부여된 각각의 트랜스코딩 작업에 대하여 네트워크 범위 및 계산 부하에 따라 해당 트랜스코딩 작업을 할당 받는다.
- [0045] 본 발명의 실시예에 따른 중앙 트랜스코딩 서버 및 엣지 서버는 각 채널이 트랜스코딩 서버의 네트워크 범위 내에 있는지 여부를 나타내는 이진 변수를 사용하여, 해당 채널의 가장 낮은 비트레이트 버전에 대한 트랜스코딩 작업 실행 여부에 따라 이진값을 부여 받는다.
- [0046] 단계(230)에서, 인기도 가중 비디오 품질에 대한 네트워크 범위 및 계산 부하에 따른 영향을 최소화하기 위하여 상기 수집 서버를 통해 일부 작업을 라이브 스트리밍 공급자가 운영하는 상기 중앙 트랜스코딩 서버에 재할당하거나 또는 제거한다.
- [0047] 단계(240)에서, 수집 서버를 통해 할당된 트랜스코딩 작업을 상기 중앙 트랜스코딩 서버 또는 상기 엣지 서버에서 수행한 후, 전송 서버를 통해 복수의 시청자 단말에 라이브 비디오 스트리밍을 제공한다.
- [0048] 본 발명의 실시예에 따른 전송 서버는 복수의 시청자 단말로부터 요청된 비트레이트를 가진 버전이 트랜스코딩 작업이 실행되지 않은 경우, 트랜스코딩 작업이 실행된 버전 중에서 상기 요청된 비트레이트보다 낮은 버전 중 가장 높은 비트레이트를 가진 버전을 전송한다.
- [0049] 이하, 제안하는 제안하는 엣지 컴퓨팅 환경에서 라이브 비디오 스트리밍 품질을 고려한 비용 최적화를 위한 트랜스코딩 작업 할당 과정에 관하여 더욱 상세히 설명한다.
- [0051] 도 3은 본 발명의 일 실시예에 따른 트랜스코딩 작업 결정 및 엣지 할당 알고리즘을 나타내는 도면이다.
- [0052] 도 4는 본 발명의 일 실시예에 따른 비용 절감을 위한 트랜스코딩 작업 재할당 및 제거 알고리즘을 나타내는 도면이다.
- [0053] N^{ch} 는 총 채널 수이다. N^{ver} 는 사용하는 각각 다른 비트레이트 버전의 수이다. $V_{i,k}$ 는 채널 i 의 k 번째 가장 낮은 비트레이트 버전 ($i=1, \dots, N^{ch}$ 및 $k=1, \dots, N^{ver}$)이다. 가장 높은 비트레이트 버전은 항상 사용할 수 있으므로 트랜스코딩으로 생성될 수 있는 버전 N^{tran} 의 수는 $N^{ver}-1$ 이다.
- [0054] $\tau_{i,k}$ 를 버전 $V_{i,k}$, ($i=1, \dots, N^{ch}$ 및 $k=1, \dots, N^{tran}$)에 대한 트랜스코딩 작업이라고 가정한다. 각 작업 $\tau_{i,k}$ 의 엣지 서버 CPU 사용량 요구 사항 $U_{i,k}$ 를 가장 강력한 ES의 CPU 프로세싱 능력에 대비한 비율로 표현한다.
- [0055] $P_{i,k}$ 를 버전 $V_{i,k}$ 의 액세스 확률, 즉 일반 시청자가 시청할 확률이라고 한다. 각 채널의 확률은 현재 조회수에서 파생될 수 있다. $Q_{i,k}$ 를 $V_{i,k}$ 버전의 비디오 품질 인덱스라고 하자. $Q_{i,k}$ 의 값은 비디오 품질 측정 도구를 사용하여 얻을 수 있다. 본 발명에서는 Netflix에서 개발한 VMAF(Video Multi-Method Assessment fusion) 지수를 사용하여 이를 표현하였다. 본 발명에서는 더 인기 있는 채널에 더 높은 품질을 제공하기 위해 $V_{i,k}$ 버전의 인기 가중치 동영상 품질(PWQ), $Q_{i,k}^{pwq}$ 을 도입하고, 아래와 같이 정의 한다:

$$Q_{i,k}^{pwq} = Q_{i,k} p_{i,k}.$$

이진 변수 $X_{i,k}$ 를 사용하여 버전 $V_{i,k}$ 의 트랜스코딩을 통한 생성 여부를 나타낸다. 즉 $\tau_{i,k}$ 가 실행되면 $X_{i,k}=1$; 그렇지 않으면 $X_{i,k}=0$ 이다.

라이브 스트리밍에서 각 시청자 단말은 해당 단말과 현재 네트워크 상황에서 지원할 수 있는 가장 높은 비트레이트 버전을 요청할 것으로 예상할 수 있다. 전달된 버전의 비트레이트가 요청된 것보다 높으면 재버퍼링을 통해 QoE를 크게 줄일 수 있으며 이로 인해 재생이 완전히 중지될 수도 있다. 어떤 경우에도 이를 방지하려면 모든 비디오를 해당 채널에서 제공하는 가장 낮은 비트레이트로 트랜스코딩해야 한다.

요청된 비트레이트를 가진 버전이 코드 변환되지 않았기 때문에 사용할 수 없는 경우, 트랜스코딩 된 버전 중에서 요청된 비트레이트보다 낮은 버전 중, 가장 높은 비트레이트를 가진 버전이 전송된다. 요청된 비트레이트가 k 인 경우 이 버전의 비디오 스트림을 $H_{i,k}$ 라고 하며 아래와 같이 정의한다:

$$H_{i,k} = \begin{cases} \arg \max_{m=1, \dots, k} X_{i,m} = 1, & \text{if } k = 1, \dots, N^{\text{tran}} \\ V_{i,N^{\text{ver}}}, & \text{if } k = N^{\text{ver}}. \end{cases}$$

$H_{i,k}$ 는 $X_{i,m}$ 값의 조합에 의해 결정되므로 채널 i 의 총 PWQ는 다음과 같이 계산할 수 있다:

$$Q_i^{\text{ch}}(X_{i,1}, \dots, X_{i,N^{\text{tran}}}) = \sum_{k=1}^{N^{\text{ver}}} Q_{i,H_{i,k}}^{pwq}.$$

다음으로, 본 발명의 실시예에 따른 엣지 서버의 사용 비용에 관하여 설명한다.

라이브 스트리밍 서비스 제공자는 엣지 서버를 사용하기 위해 비용을 지불해야 한다. 본 발명의 실시예에서는 주로 비용이 CPU 사용량에 비례하는 선형 비용 모델을 고려한다. ES j 를 사용하기 위한 비용 $C_j^{\text{cost}}(u)$ 는 정규화된 비용 값 q_i 과 CPU 사용량 u 에 비례한다. ($0 \leq u \leq U_j^{\text{max}}$ 이다) 이는 아래와 같이 표현할 수 있다:

$$C_j^{\text{cost}}(u) = \alpha_j u.$$

또한 처리량에 관계없이 주어진 기간 동안 사용 여부에 따라 엣지 서버 사용 비용이 결정되는 온-오프(on-off) 비용 모델도 고려한다. C_j^{onoff} 은 일정 기간 동안 ES j 를 사용할 때 지불하는 가장 비싼 ES의 비율로 계산된 정규화된 비용이다. 그러므로 온-오프 비용 모델에서의 $C_j^{\text{cost}}(u)$ 는 다음과 같이 결정할 수 있다:

$$C_j^{\text{cost}}(u) = \begin{cases} 0 & u = 0 \\ C_j^{\text{onoff}} & \text{otherwise.} \end{cases}$$

본 발명의 실시예에 따르면, TS j 에서 각 태스크 $\tau_{i,k}$ 가 실행되는 트랜스코딩 서버의 인덱스를 $Y_{i,k}$ 라고 하자. TS를 사용할 수 없고 $X_{i,k} = 0$ 이면 $Y_{i,k}$ 가 -1로 설정된다.

본 발명의 목표는 다음 세 가지 제약 조건에서 전체 PWQ를 최대화하는 각 작업 $\tau_{i,k}$ 에 대한 $X_{i,k}$ 및 $Y_{i,k}$ 값을 찾는 것이다: 프로세싱 용량 제약, 네트워크 범위 제약 및 네트워크 범위 제약.

본 발명의 실시예에 따른 프로세싱 용량 제약에 있어서, TS j 에서 모든 작업을 실행하는 데 필요한 총 CPU 사용량은 다음과 같이 최대 처리 용량 U_j^{max} 를 초과해서는 안 된다.

$$\sum_{\forall \tau_{i,k}, Y_{i,k}=j} U_{i,k} \leq U_j^{\text{max}}, (j = 0, \dots, N^{\text{ES}}).$$

본 발명의 실시예에 따른 네트워크 범위 제약에 있어서, 각 트랜스코딩 작업은 다음과 같이 채널 또는 중앙 트

랜스코딩 서버의 네트워크 범위 내의 엣지 서버에서 실행되어야 한다:

$$\forall \tau_{i,k}, \text{ where } X_{i,k} = 1, C_{i,Y_{i,k}}^{\text{net}} = 1.$$

본 발명의 실시예에 따른 1) 비용 예산 제약에 있어서, 발생한 총 비용은 다음과 같이 라이브 스트리밍 서비스 제공업체의 비용 예산 C^{budget} 을 초과해서는 안 된다:

$$\sum_{j=1}^{N^{\text{ES}}} C_j^{\text{cost}} \left(\sum_{\forall \tau_{i,k}, Y_{i,k}=j} U_{i,k} \right) \leq C^{\text{budget}}.$$

이제 전체 PWQ를 최대화하는 각 작업 $\tau_{i,k}$ 에 대한 $X_{i,k}$ 및 $Y_{i,k}$ 값을 찾는 문제인 트랜스코딩 작업 할당 문제 (TTAP)를 공식화할 수 있으며, 다음과 같다:

$$\text{Maximize } \sum_{i=1}^{N^{\text{ch}}} Q_i^{\text{ch}}(X_{i,1}, \dots, X_{i,N^{\text{tran}}})$$

$$\text{Subject to } \forall \tau_{i,k} \text{ where } X_{i,k} = 1, C_{i,Y_{i,k}}^{\text{net}} = 1 \\ \sum_{\forall \tau_{i,k}, Y_{i,k}=j} U_{i,k} \leq U_j^{\text{max}}, (j = 0, \dots, N^{\text{ES}}) \\ \sum_{j=1}^{N^{\text{ES}}} C_j^{\text{cost}} \left(\sum_{\forall \tau_{i,k}, Y_{i,k}=j} U_{i,k} \right) \leq C^{\text{budget}}.$$

TTAP는 NP-hard로 알려진 할당 제한(MMPAR)이 있는 객관식 다중 배낭 문제로 축소될 수 있다.

본 발명에서는 이 문제를 풀기 위해 두 단계의 휴리스틱을 도입한다. 먼저 트랜스코딩 작업 중 실행할 것을 PWQ의 최대화를 고려하여 결정하고 프로세싱 용량 제약 및 네트워크 범위 제약 내에서 트랜스코딩 서버에 할당한다. 다음, 비용 예산 제약이 충족될 때까지 PWQ 손실을 최소화하면서 첫 단계에서 할당된 작업을 하나씩 제거한다.

도 3 및 도 4를 참조하면, 첫 번째 단계는 트랜스코딩 작업 결정 및 엣지 할당 단계이다. 알고리즘 실행 동안 각 작업 $\tau_{i,k}$ 에 대해 $X_{i,k}$ 의 임시 값을 보유하는 임시 이진 변수를 $X_{i,k}^{\text{tmp}}$ 라고 한다.

변수 $D_{i,k}^{\text{task}}$ 는 실행할 작업 집합에서 $\tau_{i,k}$ 를 제거했을 때의 PWQ를 나타내며 아래와 같이 표현할 수 있다:

$$D_{i,k}^{\text{task}} = Q_i^{\text{ch}}(X_{i,1}^{\text{tmp}}, \dots, X_{i,N^{\text{tran}}}^{\text{tmp}}) - \\ Q_i^{\text{ch}}(X_{i,1}^{\text{tmp}}, \dots, X_{i,k-1}^{\text{tmp}}, 0, X_{i,k+1}^{\text{tmp}}, \dots, X_{i,N^{\text{tran}}}^{\text{tmp}}).$$

$X_{i,k}^{\text{tmp}}$ 의 모든 값은 1로 초기화되므로 $D_{i,k}^{\text{task}}$ 는 초기에는 $\tau_{i,k}$ 제거로 인한 PWQ의 변화를 나타낸다.

이제 $D_{i,k}^{\text{task}}$ 에 $U_{i,k}$ 의 비율인 $R_{i,k}^{\text{TDA}}$ 를 결정할 수 있다. 이는 아래와 같다.

$$R_{i,k}^{\text{TDA}} = \frac{D_{i,k}^{\text{task}}}{U_{i,k}}.$$

$D_{i,k}^{\text{task}}$ 의 값이 높을수록 PWQ에 미치는 영향이 크기 때문에 $\tau_{i,k}$ 를 제거하는 것은 바람직하지 않다. 따라서 $R_{i,k}^{\text{TDA}}$ 값이 더 높은 작업이 우선적으로 할당된다.

작업을 어떤 ES에 할당할 지를 결정하기 위해, 변수 P_j^{TDA} 를 통해 각 ES j에 추가 작업을 할당하는 가능성 및 타당성을 표현했다. P_j^{TDA} 의 값은 해당 ES에서 남은 프로세싱 용량과 비용을 결합한 식으로 비용 모델에 따라 다르게 정의된다.

본 발명의 실시예에 따른 선형 모델에 있어서, 비용 효율적으로 엣지 서버의 활용도를 높이려면 $D_{i,k}^{\text{task}}$ 값이 더

높은 작업을, 남은 프로세싱 용량이 많고 사용 비용이 낮은 ES에 할당해야 한다. 따라서 P_j^{TDA} 의 값은 아래와 같으며, P_j^{TDA} 값이 더 높은 ES에 우선적으로 태스크를 할당한다:

$$P_j^{TDA} = \frac{U_j^{\max} - \sum_{\forall \tau_{i,k} \in A_L} U_{i,k}}{C_j^{\text{cost}} (\sum_{\forall \tau_{i,k} \in A_L} U_{i,k})}.$$

본 발명의 실시예에 따른 온-오프(On-off) 모델에 있어서, 온-오프 비용 모델을 사용할 때 하나의 작업을 E에 할당하면 전체 고정 비용이 발생한다. 따라서 ES 간에 작업 실행을 분산시키는 것은 불리하다. 이 모델을 사용하면 ES를 최대한 활용하기 위해 P_j^{TDA} 값이 더 높은 ES에 작업을 우선적으로 할당한다:

$$P_j^{TDA} = \frac{\sum_{\forall \tau_{i,k} \in A_L} U_{i,k}}{C_j^{\text{cost}} (\sum_{\forall \tau_{i,k} \in A_L} U_{i,k})}.$$

도 3에는 트랜스코딩 작업 결정 및 엡지 할당 단계의 의사코드를 나타내었다. 가장 낮은 비트레이트 버전은 항상 트랜스코딩되어야 하므로 $\tau_{i,1} (i=1, \dots, N^{\text{ch}})$ 을 제일 먼저 ES에 할당한 다음, 다른 비트레이트 버전들을 ES에 할당한다. 알고리즘이 끝나면 TS j에 할당된 작업의 배열 A_j 를 구할 수 있다.

본 발명의 실시예에 따른 알고리즘의 비용 절감을 위한 트랜스코딩 작업 재할당 및 제거 단계에 있어서, 먼저 첫 번째 단계에서 비용 예산 제약이 충족되지 않으면 전체 PWQ를 가능한 한 적게 줄이려고 시도하면서 ES에 할당한 작업 중 일부를 제거해야 한다.

첫 번째 단계에서 ES들에 할당한 작업들의 집합을 A^{ES} 라고 하자. ($A^{\text{ES}} = A_1 \cup \dots \cup A_N^{\text{ES}}$). 이 작업이 할당된 엡지 서버의 PWQ 대 CPU 사용량의 비율은 중앙 트랜스코딩 서버에 적용되는 비율보다 클 수 있다. 이 경우 제거된 작업을 중앙 트랜스코딩 서버로 재배포할 수 있으며, 이 재배포는 비용 모델에 따라 다르게 동작한다.

본 발명의 실시예에 따른 선형 모델에 있어서, 각 ES를 사용하는 비용이 프로세서 사용량에 따라 증가하므로, 각 작업에 대한 비용 효율성을 표현하는 변수 $P_{i,k}^{\text{CR}}$ 를 다음과 같이 계산한다:

$$P_{i,k}^{\text{CR}} = \frac{D_{i,k}^{\text{task}}}{C_{I_{i,k}}^{\text{cost}} (U_{i,k})}$$

$P_{i,k}^{\text{CR}}$ 값이 가장 낮은 작업은 할당 해제 후보들을 담은 집합 A^{del} 에 추가되고 이 프로세스는 비용 예산 제약을 만족할 때까지 반복된다.

본 발명의 실시예에 따른 온-오프(On-off) 모델에 있어서, 비용 예산 제약을 만족시키기 위해서는 하나 이상의 ES에 할당된 작업을 모두 제거해야 한다. ES j에 대한 비용 효율성 변수 P_i^{CR} 는 다음과 같이 계산할 수 있다:

$$P_j^{\text{CR}} = \frac{\sum_{\forall \tau_{i,k} \in A_j, k \neq 1} D_{i,k}^{\text{task}}}{C_j^{\text{cost}} (\sum_{\forall \tau_{i,k} \in A_j} U_{i,k})}$$

P_i^{CR} 값이 가장 낮은 엡지 서버에 할당된 작업들은 집합 A^{del} 에 추가되고 이 프로세스는 비용 예산 제약을 만족할 때까지 반복된다.

마지막으로 중앙 트랜스코딩 서버에서 트랜스코딩 할 작업을 결정해야 한다. 첫 단계에서 찾은 작업 집합 A_0 와 ES에서 할당 해제된 작업 집합 A^{del} 의 합집합에서 결정된다. 이를 달성하기 위해 A^{del} 는 $A_0 \cup A^{\text{del}}$ 으로 대체되며, A_0 는 새로운 A^{del} 에 포함된 가장 낮은 비트레이트 버전의 작업들로 구성된다. 그런 다음 $Q_{i,k}^{\text{pwq}} U_{i,k}$ 값이 가장 높은 작업을 A^{del} 에서 A_0 로 반복적으로 이동하여 CTS의 용량 제한 내에서 전체 PWQ를 증가시킨다. 해당 단계가 완료

되면 A^{ES} 및 A_0 집합을 통해 $X_{i,k}$ 및 $Y_{i,k}$ 를 설정할 수 있다.

[0103] 본 발명의 실시예에 따르면, 기본적으로 N^{ch} 는 Twitch.tv에서 동시에 방송되는 중간 채널 수인 6000으로 설정되어 있다. 본 발명에서는 공용 트랜스코딩 클라우드 제공업체인 Zencoder에서 제공하는 비트레이트 세트를 사용하여, 각 해상도에 따라 아래의 표 1에서 a와 b 값과 생성할 버전의 비트레이트, r(Mbps 단위)을 사용하여 회귀 모델 계산식, arb에 대입하여 각 작업의 CPU 사용량을 구했다.

[0104] 표 1은 라이브 스트리밍에서 1080p 2,750kbps 버전을 다른 비트레이트 버전으로 트랜스 코딩하는 데 필요한 CPU 주파수(GHZ)를 결정하는 파라미터이다.

[0105] <표 1>

Target resolution	a	b
224p	0.673091	0.024642
360p	0.827912	0.033306
720p	1.341512	0.060222
1080p	1.547002	0.080571

[0106]

[0107] VMAF 값은 정규 분포로 모델링되며 각 비트레이트에서 버전에 대한 VMAF 값은 측정값을 통해 얻어지며, 각 채널의 액세스 확률을 감마 분포로 모델링했으며, k와 θ 값은 Twitch.tv에서 얻은 값인 각각 0.399와 14260로 설정했다.

[0108] 서로 다른 버전의 상대 인기도를 표현하기 위해, μ 의 평균값과 σ 의 분산을 갖는 정규분포를 사용했다. σ 값은 독립적으로 설정되며, 각 사례에서 가장 인기 버전의 우세를 결정한다. 본 발명에서는 중간 비트일수록 인기도가 높을 경우를 디폴트로 하여, σ 를 1, $\mu=Nver/2$ 를 사용했다.

[0109] 엣지 서버의 위치를 결정하기 위해 호주 멜버른의 실제 셀룰러 네트워크 위치 데이터를 사용하였다. 또한 엣지 서버의 CPU 특성은 2020년 또는 2021년에 출시된 Asus RS620SA-E10-RS12, Dell PowerEdge R7525, HPE Apollo XL225n Gen10 Plus, Fujitsu Primergy RX4770 M6, Lenovo Technology ThinkSystem SR665를 기반으로 했다.

[0110] 기본적으로 선형 비용 모델을 사용한다. Amazon Web Services에서 운영하는 컴퓨팅 자원과 관련된 6가지 비용 모델(c4, c5n, c5a, c5, c6gn 및 c6g)에서 정규화된 비용 값 q_i 를 얻었으며 이는 0.63에서 1 사이의 값이었다. 본 발명에서는 0.63에서 1 사이에서 랜덤으로 q_i 를 선택했다.

[0111] 본 발명에서는 ES 할당 및 작업 선택의 6개의 벤치마크 조합과 제안하는 트랜스코딩 작업 할당 기법을 비교했다. 6개의 벤치마크 조합은 각각 RA(무작위 ES 할당)+AP(모든 인기 채널의 버전 트랜스코딩), RA+HPF(가장 인기 있는 태스크 우선), LUF(가장 낮은 사용률의 ES 우선)+AP, LUF+HPF, LCF (최소비용의 ES 우선)+AP 및 LCF+HPF이다. 본 발명에서 제안하는 트랜스코딩 작업 할당 기법의 알고리즘을 TDA+CR라고 한다.

[0112] 도 5는 본 발명의 일 실시예에 따른 트랜스코딩 작업 할당 기법과 종래기술의 기법과의 비용 제한에 따른 인기도 가중 품질을 비교한 도면이다.

[0113] N^{ch} 가 6000이고 N^{ES} 가 100일 때 비용 예산이 총 PWQ에 미치는 영향을 조사했다. 이를 위해 C^{ratio} 변수를 도입했다. 이는 ES j를 최대 사용률 U_j^{max} 로 사용하는 비용에 대한 비용 예산의 비율이며 다음과 같다:

$$C^{ratio} = \frac{C^{budget}}{\sum_{j=1}^{N^{ES}} C_j^{cost}(U_j^{max})}.$$

- [0114] 본 발명에서 제안하는 기법인TDA+CR은 항상 최고의 성능을 달성하여 다른 방법보다 1.06배에서 2.13배(평균 1.45배) 더 많은 PWQ를 생성한다. 차이는 C^{ratio} 가 감소함에 따라 증가하는 경향이 있으며, 이는 제안하는 기법이 감소된 비용 예산에 대처하는 데 효과적임을 나타낸다.
- [0116] 도 6은 본 발명의 일 실시예에 따른 트랜스코딩 작업 할당 기법과 종래기술의 기법과의 채널 수에 따른 인기도 가중 품질리티를 비교한 도면이다.
- [0117] C^{ratio} 가 50%, N^{ES} 가 100일 때 비용 예산이 총 PWQ에 미치는 영향을 조사했다. 제안하는 기법인 TDA+CR은 다른 방법보다 1.08배에서 1.97배(평균 1.44배) 더 많은 PWQ를 달성한다.
- [0118] 특히 벤치마크 방식 중 최고인 LCF+HPF보다 1.08~1.19배(평균 1.14배) 높은 PWQ를 달성한다.
- [0119] 채널 수가 증가함에 따라, 높은 비트 전송률 버전이 더 적게 트랜스코딩되므로 PWQ가 떨어진다. 이 떨어지는 정도는 비교 벤치마크 계획보다 제안하는 기법이 훨씬 적으므로 더 우수하다.
- [0121] 도 7은 본 발명의 일 실시예에 따른 트랜스코딩 작업 할당 기법과 종래기술의 기법과의 엣지 서버 수에 따른 인기도 가중 품질리티를 비교한 도면이다.
- [0122] 예상대로 총 PWQ는 엣지 서버의 개수의 증가에 따라 함께 증가한다. 트랜스코딩에 더 많은 프로세싱 용량을 사용할 수 있기 때문이다.
- [0123] 제안하는 기법은 다른 방법보다 1.05배에서 2.03배(평균 1.45배) 더 많은 PWQ를 얻는다. 이 차이는 엣지 서버가 적을수록 더 커지며, 이는 엣지 서버의 총 프로세싱 용량이 낮을 때 타 기법보다 제안하는 기법이 더 효과적으로 작동함을 보인다.
- [0125] 도 8은 본 발명의 일 실시예에 따른 트랜스코딩 작업 할당 기법과 종래기술의 기법과의 비용 제한에 따른 온-오프(On-off) 비용 모델에서의 인기도 가중 품질리티를 비교한 도면이다.
- [0126] 지금까지 제시된 모든 결과는 선형 비용 모델을 사용하여 얻어졌다. 그러므로 온-오프(On-off) 비용 모델에 대한 실험을 수행했다. 선형 비용 모델과의 공정한 비교를 위해 C_j^{onoff} 의 각 값을 $\alpha_j U_j^{max}$ 로 설정했다.
- [0127] 또한 LCF 방법은 현재 비용 사용을 기반으로 다음 작업을 할당하기 때문에 이 모델에서 변하지 않는 온-오프 비용 모델에 직접 사용할 수 없다. 따라서 본 발명에서는 가장 저렴한 ES를 선택하고 ES가 완전히 활용될 때까지 작업을 할당하는 LCF-ONOFF를 도입했다.
- [0129] 도 9는 본 발명의 일 실시예에 따른 트랜스코딩 작업 할당 기법의 인기도 가중 품질리티를 선형 비용 모델과 온-오프(On-off) 비용 모델에서 비교한 도면이다.
- [0130] 온-오프 비용 모델을 사용할 때 N^{ch} 가 6000이고 N^{ES} 가 100일 때 비용 예산이 총 PWQ에 미치는 영향을 조사했다. 제안하는 기법은 온-오프 비용 모델에서도 최고의 성능을 달성하여 다른 방법보다 1.06배에서 1.98배(평균 1.34배) 더 많은 PWQ를 얻는다. 비용 예산이 낮을 때 이 격차는 더 두드러진다.
- [0131] 선형 비용 모델은 항상 온-오프 비용 모델보다 PWQ가 더 높다. 이 차이는 0.24%와 6.63%(평균 2.34%) 사이이며 C^{ratio} 가 감소함에 따라 증가한다. 이는 온-오프 비용 모델이 상대적으로 적은 수의 ES를 사용하도록 강제하는 반면 선형 비용 모델은 특정 작업에 가장 적합한 ES를 더 자주 할당하도록 허용하기 때문이다. 이러한 실험 결과들을 통해 제안한 기법이 언제나 우수함을 확인했다.
- [0133] 이상에서 설명된 장치는 하드웨어 구성요소, 소프트웨어 구성요소, 및/또는 하드웨어 구성요소 및 소프트웨어 구성요소의 조합으로 구현될 수 있다. 예를 들어, 실시예들에서 설명된 장치 및 구성요소는, 예를 들어, 프로

세서, 콘트롤러, ALU(arithmetic logic unit), 디지털 신호 프로세서(digital signal processor), 마이크로컴퓨터, FPA(field programmable array), PLU(programmable logic unit), 마이크로프로세서, 또는 명령(instruction)을 실행하고 응답할 수 있는 다른 어떠한 장치와 같이, 하나 이상의 범용 컴퓨터 또는 특수 목적 컴퓨터를 이용하여 구현될 수 있다. 처리 장치는 운영 체제(OS) 및 상기 운영 체제 상에서 수행되는 하나 이상의 소프트웨어 애플리케이션을 수행할 수 있다. 또한, 처리 장치는 소프트웨어의 실행에 응답하여, 데이터를 접근, 저장, 조작, 처리 및 생성할 수도 있다. 이해의 편의를 위하여, 처리 장치는 하나가 사용되는 것으로 설명된 경우도 있지만, 해당 기술분야에서 통상의 지식을 가진 자는, 처리 장치가 복수 개의 처리 요소(processing element) 및/또는 복수 유형의 처리 요소를 포함할 수 있음을 알 수 있다. 예를 들어, 처리 장치는 복수 개의 프로세서 또는 하나의 프로세서 및 하나의 콘트롤러를 포함할 수 있다. 또한, 병렬 프로세서(parallel processor)와 같은, 다른 처리 구성(processing configuration)도 가능하다.

[0134] 소프트웨어는 컴퓨터 프로그램(computer program), 코드(code), 명령(instruction), 또는 이들 중 하나 이상의 조합을 포함할 수 있으며, 원하는 대로 동작하도록 처리 장치를 구성하거나 독립적으로 또는 결합적으로(collectively) 처리 장치를 명령할 수 있다. 소프트웨어 및/또는 데이터는, 처리 장치에 의하여 해석되거나 처리 장치에 명령 또는 데이터를 제공하기 위하여, 어떤 유형의 기계, 구성요소(component), 물리적 장치, 가상 장치(virtual equipment), 컴퓨터 저장 매체 또는 장치에 구체화(embodiment)될 수 있다. 소프트웨어는 네트워크로 연결된 컴퓨터 시스템 상에 분산되어서, 분산된 방법으로 저장되거나 실행될 수도 있다. 소프트웨어 및 데이터는 하나 이상의 컴퓨터 판독 가능 기록 매체에 저장될 수 있다.

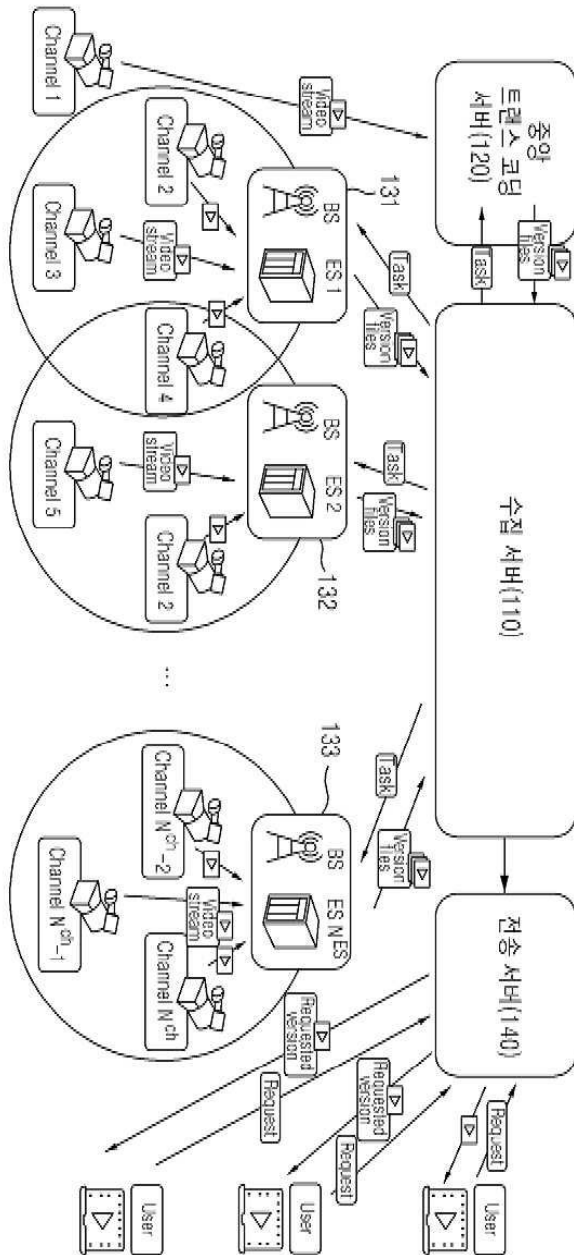
[0135] 실시예에 따른 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 상기 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 상기 매체에 기록되는 프로그램 명령은 실시예를 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다. 컴퓨터 판독 가능 기록 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(magnetic media), CD-ROM, DVD와 같은 광기록 매체(optical media), 플롭티컬 디스크(floptical disk)와 같은 자기-광 매체(magneto-optical media), 및 롬(ROM), 램(RAM), 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다. 프로그램 명령의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다.

[0136] 이상과 같이 실시예들이 비록 한정된 실시예와 도면에 의해 설명되었으나, 해당 기술분야에서 통상의 지식을 가진 자라면 상기의 기재로부터 다양한 수정 및 변형이 가능하다. 예를 들어, 설명된 기술들이 설명된 방법과 다른 순서로 수행되거나, 및/또는 설명된 시스템, 구조, 장치, 회로 등의 구성요소들이 설명된 방법과 다른 형태로 결합 또는 조합되거나, 다른 구성요소 또는 균등물에 의하여 대치되거나 치환되더라도 적절한 결과가 달성될 수 있다.

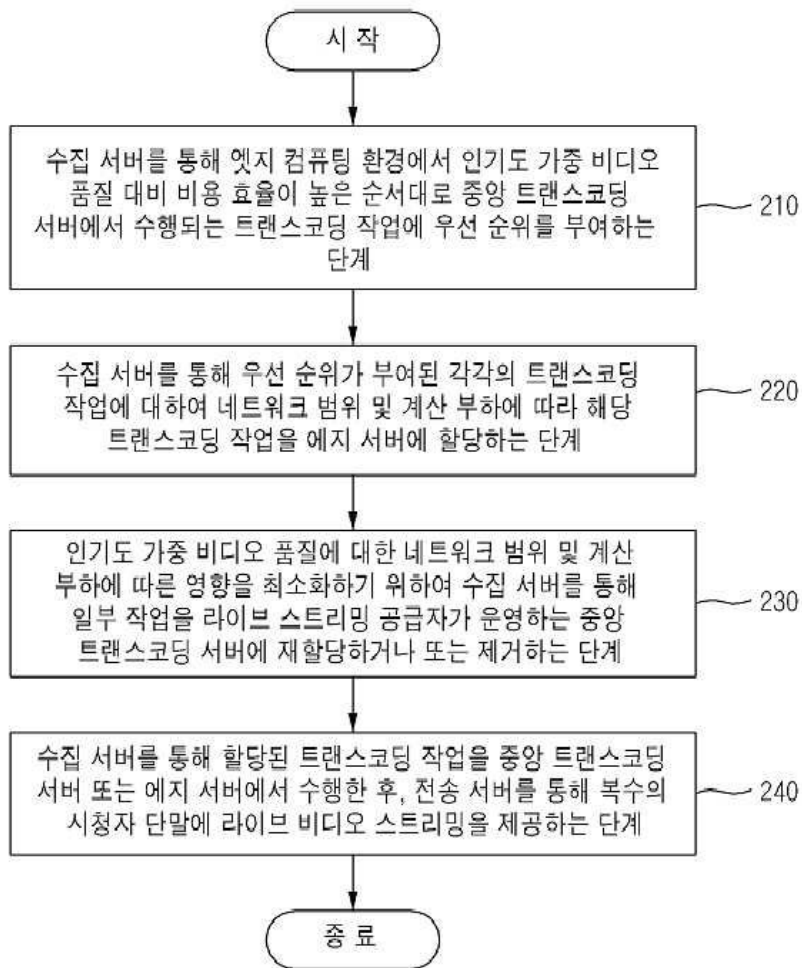
[0137] 그러므로, 다른 구현들, 다른 실시예들 및 특허청구범위와 균등한 것들도 후술하는 특허청구범위의 범위에 속한다.

도면

도면1



도면2



도면3

Algorithm 1: Task determination and allocation (TDA) phase.

Input: S^R : Set of $R_{i,k}^{TDA}$ values;

Output: A_j , ($j = 0, \dots, N^{ES}$);

```

1 Temporary variables:  $X_{i,k}^{tmp}$ ,
  ( $i = 1, \dots, N^{ch}$  and  $k = 1, \dots, N^{tran}$ ) ;
2 All the values of  $X_{i,k}^{tmp}$  are initialized to 1 ;
3 while  $S^R \neq \phi$  do
4   if  $\exists i, R_{i,1}^{TDA} \in S^R$  then
5     Find the task  $\tau_{i,1}$  with the highest value of
       $R_{i,1}^{TDA} \in S^R$  for which  $i = W$  and  $k = M$ ,
      and remove  $R_{W,M}^{TDA}$  from  $S^R$ ;
6   end
7   else
8     Find the task  $\tau_{i,k}$ , ( $k \neq 1$ ) with the highest
      value of  $R_{i,1}^{TDA} \in S^R$  for which  $i = W$  and
       $k = M$ , and remove  $R_{W,M}^{TDA}$  from  $S^R$ ;
9   end
10  Find the ES index  $j$ , ( $j = 1, \dots, N^{ES}$ ) with the
      highest value of  $P_j^{TDA}$  where
       $\sum_{\tau_{i,k} \in A_j} U_{i,k} + U_{W,M} \leq U_j^{max}$  and  $C_{W,j}^{net} = 1$ ,
      for which  $j = L$  ;
11  if ES  $L$  exists then
12     $A_L = A_L \cup \{\tau_{W,M}\}$ ;
13  end
14  else if  $\sum_{\tau_{i,k} \in A_0} U_{i,k} + U_{W,M} \leq U_0^{max}$  then
15     $A_0 = A_0 \cup \{\tau_{W,M}\}$ ;
16  end
17  else
18     $X_{i,k}^{tmp} \leftarrow 0$ ;
19  end
20 end

```

도면4

Algorithm 2: Cost reduction (CR) phase.

Input: C^{budget}, A_j ($j = 0, \dots, N^{\text{ES}}$), $I_{i,k}^{\text{ES}}, V_{i,k}$ and $D_{i,k}^{\text{task}}$ ($i = 1, \dots, N^{\text{ch}}$ and $k = 1, \dots, N^{\text{tran}}$);

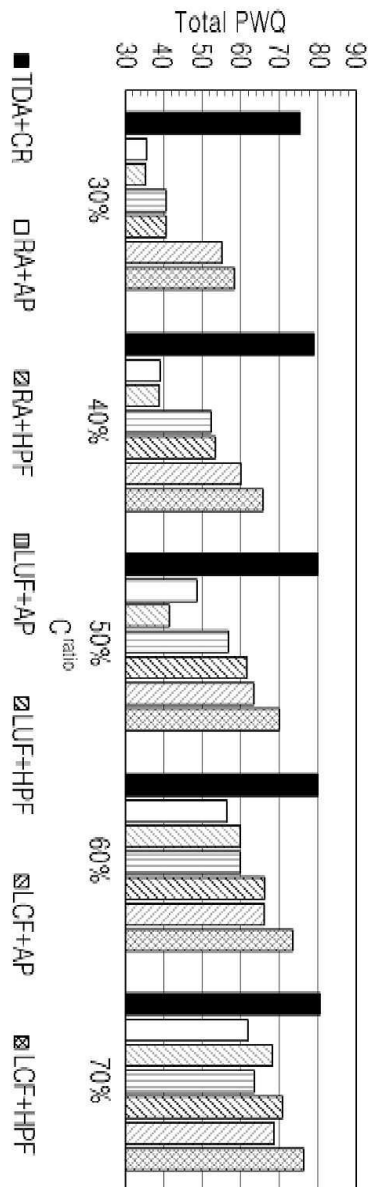
Output: $X_{i,k}$ and $Y_{i,k}$, ($i = 1, \dots, N^{\text{ch}}$ and $k = 1, \dots, N^{\text{tran}}$);

```

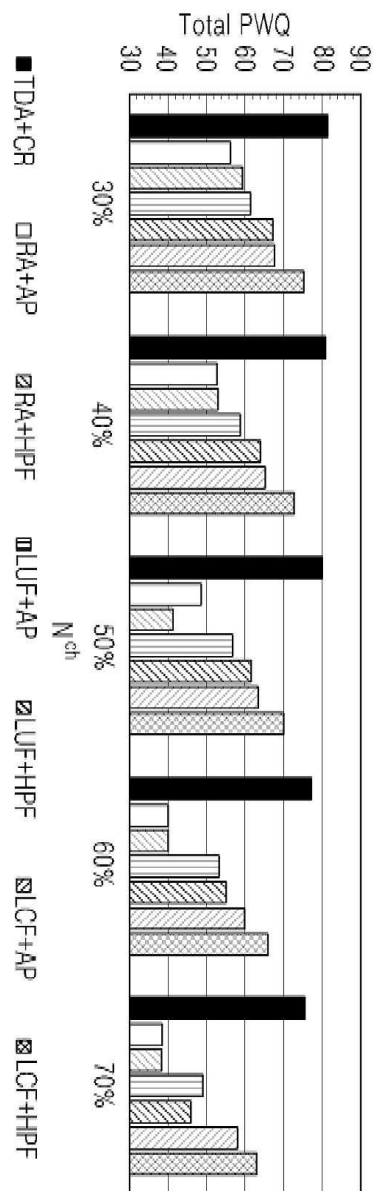
1 Sets of tasks:  $A^{\text{del}}$  and  $A^{\text{ES}}$ , ( $A^{\text{del}} \leftarrow \phi$  and  $A^{\text{ES}} \leftarrow A_1 \cup \dots \cup A_{N^{\text{ES}}}$ );
2 /* Task removal step */;
3 if Linear cost model is used then
4   while  $A^{\text{ES}} \neq \phi$  do
5     Find the lowest value of  $P_{i,k}^{\text{CR}}$  for which  $i = L$ 
        and  $k = M$ ;
6      $A_{L,M} \leftarrow A_{L,M} - \{\tau_{L,M}\}$ ;
7      $A^{\text{ES}} \leftarrow A^{\text{ES}} - \{\tau_{L,M}\}$ ;
8      $A^{\text{del}} \leftarrow A^{\text{del}} \cup \{\tau_{L,M}\}$ ;
9     if  $\sum_{j=1}^{N^{\text{ES}}} C_j^{\text{cost}}(\sum_{\tau_{i,k} \in A_j} U_{i,k}) \leq C^{\text{budget}}$ 
        then
10      Break the loop;
11    end
12  end
13 end
14 if On-off cost model is used then
15   while  $A^{\text{ES}} \neq \phi$  do
16     Find the ES index  $j$  with the lowest value of
         $P_j^{\text{CR}}$  for which  $j = L$ ;
17      $A_L \leftarrow \phi$ ;
18      $A^{\text{ES}} \leftarrow A^{\text{ES}} - A_L$ ;
19      $A^{\text{del}} \leftarrow A^{\text{del}} \cup A_L$ ;
20     if  $\sum_{j=1}^{N^{\text{ES}}} C_j^{\text{cost}}(\sum_{\tau_{i,k} \in A_j} U_{i,k}) \leq C^{\text{budget}}$ 
        then
21      Break the loop;
22    end
23  end
24 end
25 /* Task redistribution step */;
26  $A^{\text{del}} \leftarrow A^{\text{del}} \cup A_0$ ;
27  $A_0 \leftarrow \{\tau_{i,1} | \forall \tau_{i,1} \in A^{\text{del}}\}$ ;
28 while  $\sum_{\tau_{i,k} \in A_0} U_{i,k} \leq U_0^{\text{max}}$  do
29   Find a task  $\tau_{i,k} \in A^{\text{union}}$  with highest value of
         $\frac{Q_{i,k}^{\text{res}}}{U_{i,k}}$  for which  $i = L$  and  $k = M$ ;
30    $A_0 \leftarrow A_0 \cup \{\tau_{L,M}\}$ ;
31 end
32 /* Finalization */;
33 for  $i = 1$  to  $N^{\text{ch}}$  do
34   for  $k = 1$  to  $N^{\text{ver}} - 1$  do
35     if  $\tau_{i,k} \in A^{\text{ES}} \cup A_0$  then
36        $X_{i,k} \leftarrow 1$ ;
37       if  $\tau_{i,k} \in A_j$  then
38          $Y_{i,k} \leftarrow j$ ;
39       end
40     else
41        $X_{i,k} \leftarrow 0$ ;
42        $Y_{i,k} \leftarrow -1$ ;
43     end
44   end
45 end

```

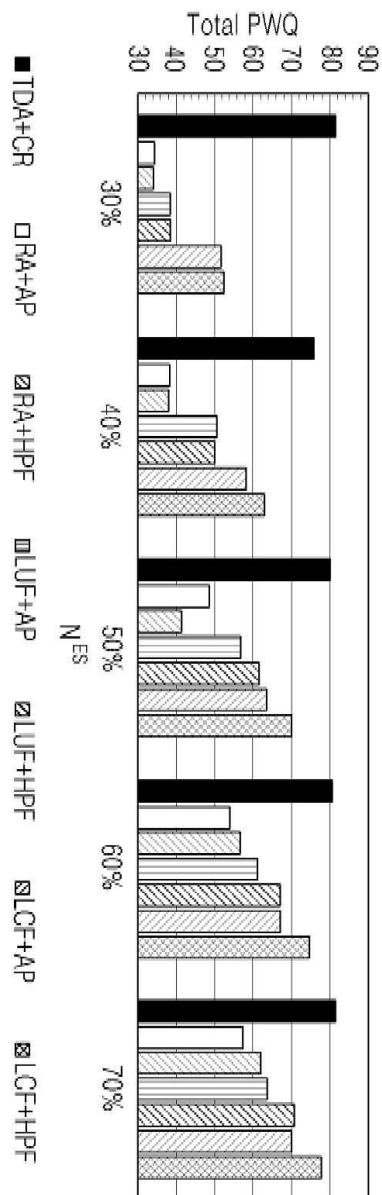
도면5



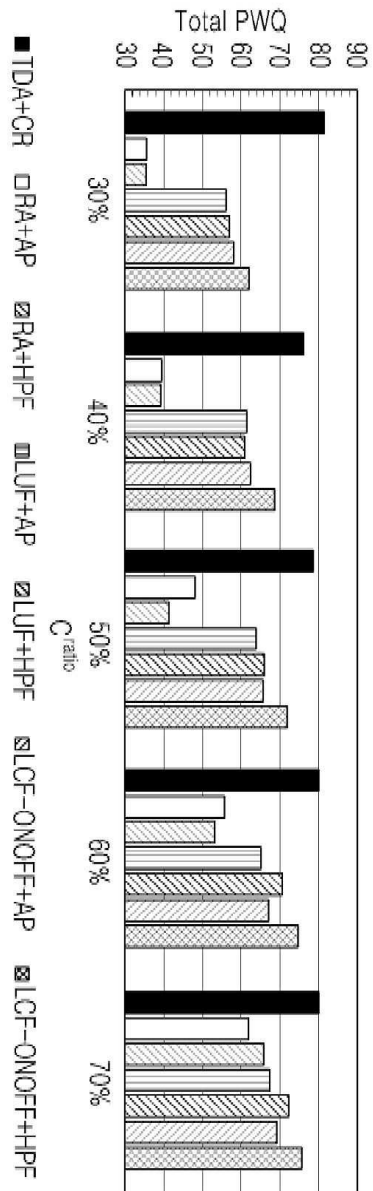
도면6



도면7



도면8



도면9

